Identifying Representative Reviewers in Internet Social Media*

Sang-Min Choi[1], Jeong-Won Cha[2], and Yo-Sub Han[1,]**

[1] Department of Computer Science, Yonsei University, Seoul, Republic of Korea
{jerassi,emmous}@cs.yonsei.ac.kr

[2] Department of Computer Engineering, Changwon National University, Changwon, Republic of Korea
jcha@changwon.ac.kr

Abstract. In a society, we have many forms of relations with other people from home, work or school. These relationships give rise to a social network. People in a social network receive, provide and pass lots of information. We often observe that there are a group of people who have high influence to other people. We call these high influence people opinion leaders. Thus, it is important and useful to identify opinion leaders in a social network. In Web 2.0, there are many user participations and we can create a social network from the user activities. We propose a simple yet reliable algorithm that finds opinion leaders in a cyber social network. We consider a social network of users who rate musics and identify representative users of the social network. Then, we verify the correctness of the proposed algorithm by the T-test.

Keywords: Social network, opinion leaders, representative reviewers, music.

1 Introduction

There are lots of information from mass media and the information affects us in various ways. However, one interesting thing is that most people often receive information not by mass media directly but by opinions of some other persons. We call these persons who influence other people *opinion leaders*. Note that opinion leaders affect people more than mass media [8]. The general public does not accept information by mass media uncritically whereas they accept information from opinion leaders easily. This implies that opinion leaders are representatives of a social network. Most people use the Internet and, thus, the Web becomes a place to share information for the general public. Since there are many users

* Choi and Han were supported by the Basic Science Research Program through NRF funded by MEST and Cha was supported by the IT R&D program of MKE/IITA 2008-S-024-01.

** Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part II, LNAI 6422, pp. 22–30, 2010.

and lots of user participations, the Web become a cyber society. The current Web shows that lots of information creates a community among peoples regardless of age, gender or nationality. Moreover, there are some influential people in the community. There are many researches to find opinion leaders in a Web social network [2,9]. We design an algorithm that finds opinion leaders in a social network of users for musics. We apply the proposed algorithm in Yahoo! music data [1] and verify the method using the T-test.

In Section 2, we revisit the previous approaches to identify opinion leaders in Twitter[1] and blogs. Then, we propose a new algorithm that finds opinion leaders in an Internet social network in Section 3. Then, we verify the algorithm using the T-test in Section 4. We conclude the paper with future directions of this research in Section 5.

2 Related Work

We briefly describe previous researches that find opinion leaders in Twitter and blogs.

2.1 Twitter

Twitter is one of the most popular social networking applications for Internet users [4]. Twitter composes a social network using a function called follow that allows users to add another users in the user lists. It also provides other functions such as reply and post. The research to find opinion leaders in Twitter is based on functions in Twitter such as follower or retweet.

The Number of Followers: Follower means the users following a particular user called followee. Followers can confirm uploaded posts by followee, and reply or retransmit (called retweet) it immediately. Namely, a follower is a user who can immediately response behavior of followee. The right side of Table 1 shows the followee who has the most followees in top 15.

The Number of Retweets: Tweet is a post of Twitter users. Followers can confirm tweet of followee and also deliver tweet to follower of their own. In this delivery case, followers can add their message.

Twitter users response their opinion for others tweets using different styles and ways [5]. In Fig. 1, users who have incoming edges are followees, and users who have outgoing edges are followers. For instance, 1 denotes a followee of users 2, 3, . . . , 8. When 1 uploads a message (tweet) to its own twitter, all users from 2 to 8 can read the message immediately. Furthermore, user 8 can retweet the message to users 9, 10 and 11 or response the message by adding opinions. This function is called retweet. Namely, a tweet that has many retweets can be considered as an influential opinion in the Twitter network and a followee

[1] `http://twitter.com`

Table 1. Number of Followers VS. Number of Retweets in Twitter [9]

	most number of retweets		most number of followers	
Rank	Name	Remark	Name	Remark
1	Pete Cashmore	News on social media	Ashton Kutcher	Actor
2	BNO News	News	Britney Spears	Musician
3	TweetMeme	News on Twitter	Ellen DeGeneres	Show host
4	Oxfordgirl	Journalist	CNN Breaking News	New
5	CNN Breaking News	News	Oprah Winfrey	Show host
6	Michael Arrington	News on technology	Twitter	Twitter
7	Fabolous	Musician	Barack Obama	President of U.S.
8	The New York Times	News	Ryan Seacrest	Show host
9	Lil duval	Comedian	THE REAL SHAQ	Sport star
10	Iran	about Iran	Kim Kardashian	Model
11	ESPN Sports News	News	John Mayer	Musician
12	Persiankiwi	about Iran	Demi Moore	Actress
13	Ashton Kutcher	Actor	Iamdiddy	Musician
14	Raymond Jahan	about Iran	Jimmy Fallon	Actor
15	Alyssa Milano	actress	Lance Armstrong	Sport star

who has many retweets can be considered as an opinion leader. The left side of Table 1 shows the top 15 users who have the most number of retweets.

The left side of the top 15 are usually celebrities in Table 1. However, we cannot infer whether followers of the celebrities react as faithful fans or having other rational reasons. Because of these reasons, we cannot conclude that the celebrities become the opinion leaders in Tweeter. Whereas, the mass media appear in the left side of Table 1. This difference show that followers react more sensitively in tweet from mass media that celebrities. In conclusion, we can consider the group of followers in the left side of Table 1 opinion leaders.

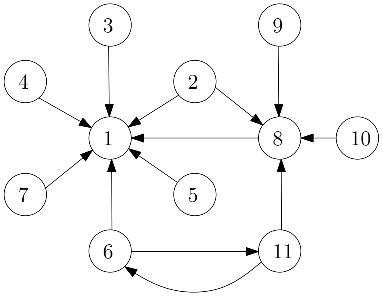

Fig. 1. An example of social network construction by the function following in Twitter [13]. The directed edge shows the following relation between users.

2.2 Blogs

Blog is an application that supplies various information and functions in the Web. One of the important functions of blog is to post various media contents and to tag the relevant information among each other. The blog influences the user in the Web. For example, the preferences of some influential bloggers affect the consume of visitors [7,10]. Thus, many advertisers can make big profits by advertising in the influential blog [12]. Hence, the influential bloggers can be opinion leaders. Some experiments examine essential issues of identifying influential bloggers, evaluate the effects of various collectible statistics from a blog site on determining blog-post influence, develop unique experiments using Digg[2] and conduct experiments by using the whole history of blog posts.

The research to find influential bloggers classifies the characteristics of bloggers into active, inactive, influential and non-influential bloggers based on intuition which active bloggers not equal to influential bloggers. The active bloggers mean users who often list their posts, and influential bloggers mean users who have influential posts. This research determines the influential posts using social gestures such as comments, incoming links, outgoing links and length of posts, and find influential bloggers based on characters of bloggers.

This research shows us the clear conclusion for the existence of influential bloggers, and how much it relates with each other between the influential bloggers and other common visitors [2].

3 Proposed Method

We identify the representative user of evaluative group of music using Yahoo! music data [1]. And we test the extracted users to know that users can be representative like a opinion leader. The average rating of music means collective opinion. This value can consider with the representative figure from collective opinion. We use the T-test to verify whether or not identified users are propose representative of the group.

In the case of adding new song, because rating of opinion leaders for new songs has representativeness, the opinion leaders can process these songs. For example, the searching or recommendation systems exclude the new songs because of a lack of information. In that case, ratings of opinion leaders have representative like a average rating. Rating of opinion leader become one of the way to solve cold start problem in information search and recommendation system [3,11].

3.1 Data Set

We use Yahoo! Webscope music data. The data is as follows:

1. User id: There exist 15,400 users in total, and given by integer number form 1 to 15,400.

[2] `http://www.digg.com`

2. Song id: There exist 1,000 songs in total, and given by integer number form 1 to 1,000.
3. Rating: As integer number from 1 to 5 there exist approximately 300,000 ratings.

In the dataset, each user gives rating to at least 10 songs. The rating presents the number from 1 to 5. The higher number means the higher rating.

3.2 Identifying Representative Reviewer

We use the following equation for identifying representative users from in the music rater group:

$$U_s = \frac{\sum_{i \in A} |(R_S(i) - R_\mu(i)|}{|A|}, \tag{1}$$

where

- A is a set of songs evaluated by user S.
- $|A|$ is the cardinality of A.
- $R_S(i)$ is the rating of song i by user S.
- $R_\mu(i)$ is the average rating of song i in the database.

Remark that the result of Equation (1) shows how close each rating of user compared to the average rating. We select those who have low score from Equation (1) as opinion leaders.

Table 2. Top 10 scores and bottom 10 scores out of Equation (1)

Rank	score	Rank	score
1	0.2598	1	2.6126
2	0.2989	2	2.5453
3	0.2996	3	2.5056
4	0.3047	4	2.5015
5	0.3086	5	2.4922
6	0.3126	6	2.4856
7	0.3140	7	2.4660
8	0.3211	8	2.4087
9	0.3219	9	2.3977
10	0.3277	10	2.3888

Table 2 shows the top 10 users and the bottom 10 users by Equation (1). For instance, the top ranked user has a gap of 0.2598 between its own rating and the average rating in Table 2 whereas the bottom ranked user has a gap of 2.6126.

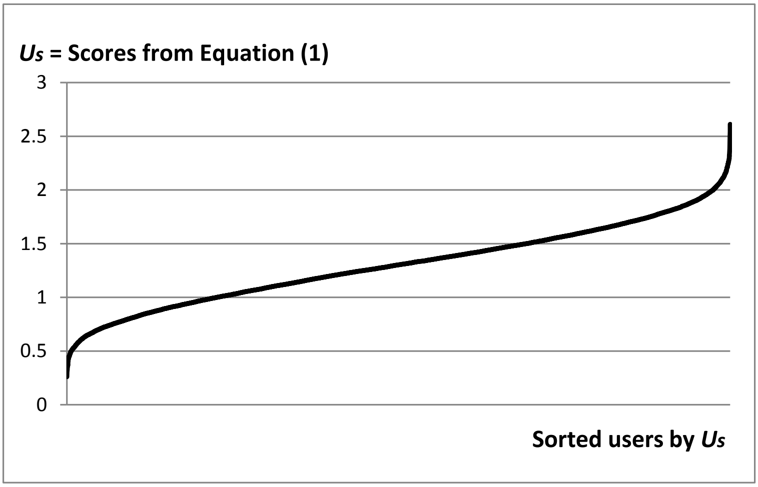

Fig. 2. The distribution of scores (U_s) from Equation (1) in ascending order

In Fig. 2, the y-axis is the value of Equation (1) and the x-axis is the set of users that are sorted by the score of Equation (1) in ascending order.

Fig. 3 illustrates the procedure of identifying representative reviewer in music rater group. First, we extract all users who evaluate songs at least 10 times. Second, applying Equation (1) to the user group. We sort the results of Equation (1) in ascending order, and select high rank users as representative reviewers in music rater group.

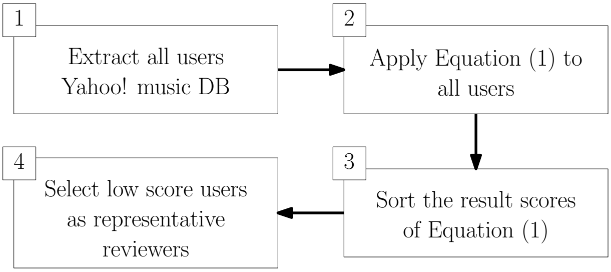
1
Extract all users Yahoo! music DB
2
Apply Equation (1) to all users
3
Sort the result scores of Equation (1)
4
Select low score users as representative reviewers

Fig. 3. The procedure of identifying representative users from Yahoo! music dataset

4 Experiments and Analysis

Now we demonstrate the validness of our algorithm for identifying representative users by the T-test. The T-test is a statistical hypothesis test in which the test statistic follows a certain distribution if the null hypothesis is supported. For details on the T-test, refer to the text [6]. We compare two sets using the T-test: The first is a set of ratings and the second is a set of average ratings. We know that rating of user can represent an average rating. Note that our null hypothesis is valid when the result of T-test (called *P-value*) is more than significance level, and therefore, those users are representativeness. Alternative hypothesis is the opposite of the null hypothesis.

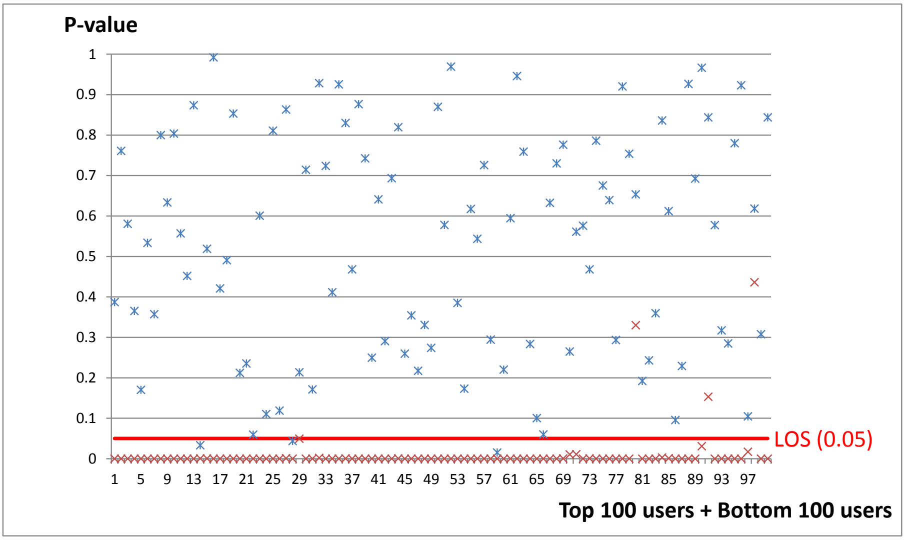

Fig. 4. Result of the T-test with level of significance (LOS) 0.05 depicted by the red line. The * denotes the top 100 users and the x denotes the bottom 100 users.

Fig. 4 shows a result of T-test applying to the 100 high ranking users and the 100 low ranking users from Equation (1). The X-axis is users and the y-axis is p-value from the T-test. In Fig. 4, for example, the P-value of the first user from the top 100 is 0.388 and the P-value of the first user from the bottom 100 is very close to 0. Since the level of significance is 0.05, we can say that the first user from the top 100 has representativeness. On the other hand, the first user from the bottom 100 does not have representativeness because the P-value is smaller than the significance level. In Fig. 4, when the significance level is 0.05, all users from the top 100 except for three users are above the significance level and, thus, these users are representative reviewers. On the other hand, we say that the bottom 100 users are not representative reviewers since their P-values are less than 0.05.

Fig. 5 shows the top 100 and the bottom 100 users when the significance level is 0.01. The significance level 0.01 contains more users than 0.05. In the significance level 0.01 and 0.05, we cannot reject null hypothesis for most top 100 users. We can conclude that most top 100 users have representativeness in

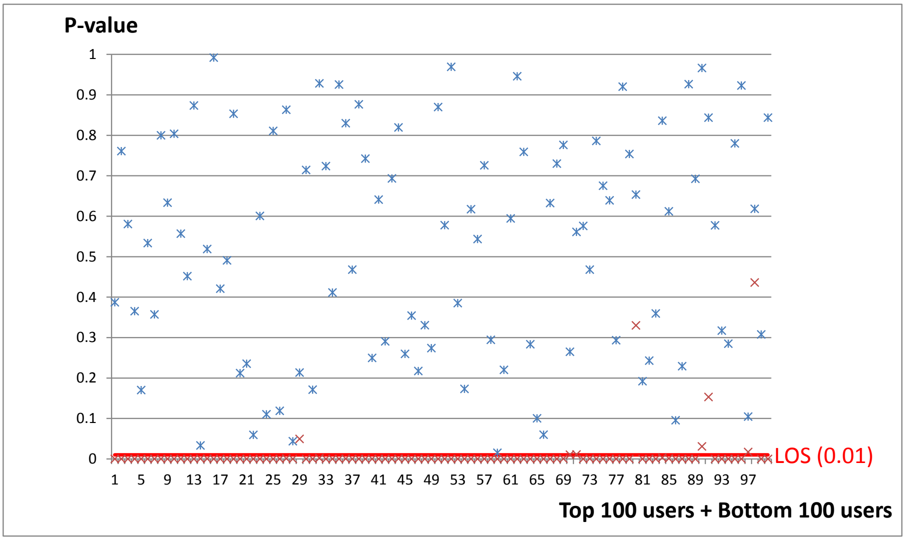

Fig. 5. Result of the T-test with level of significance (LOS) 0.01 depicted by the red line. The * denotes the top 100 users and the x denotes the bottom 100 users.

their music rater group. Most of the bottom 100 users are, on the other hand, in reject position of null hypothesis. Thus, we can say that the extracted users through our approach have representativeness.

5 Conclusions and Future Work

There are many different types of user participations in Web 2.0 and we can construct a social network of users based on these participations. For example, in Twitter, users use a special function called follow to other users. This function gives a network of users. Similarly, in blogs, people post an article and other users response it by comments or trackback. Again, we can make a social network of blog users based on these activities. In a real world society, people often make an opinion based on their trustful opinion leaders rather than mass media directly. An opinion leader is a person who has high influence to other people in a community. Since, we now have a social network in Web, it is natural to identify opinion leaders from the network. As a special opinion leader, we consider a user who has an accurate rating to items among many users; we call such user *representative user*. We have designed an algorithm that identifies representative users among many users based on the history of their ratings. We, then, have applied the proposed algorithm to the Yahoo! music dataset and demonstrated the usefulness of the algorithm by the T-test: We have formulated a hypothesis that opinions of influential users can represent all users and verified the statement.

Given overwhelming amount of information in Web, it is not easy to search good information solely by machines. Moreover, there are several user participation in Web 2.0. Researchers investigate how to use user participations for finding good information. One of the most well-known examples is collaborative filtering. There are many applications of collaborative filtering such as recommendation systems in Amazon[3] or Simania[4], information retrieval or decision support systems in Google News[5]. However, we must have enough data to find good information. Namely, we cannot use the collaborative filtering without sufficient user data [3,11]. For resolving this problem, a method of finding representative users or opinion leaders is very useful. We plan to apply the proposed algorithm to a recommendation system with small size of users for movies or songs.

References

1. Webscope from yahoo! labs, `http://webscope.sandbox.yahoo.com`
2. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM, pp. 207–218 (2008)
3. Billsus, D., Pazzani, M.J.: Learning collaborative information filters. In: ICML, pp. 46–54 (1998)
4. Blake, B.P., Agarwal, N., Wigand, R.T., Wood, J.D.: Twitter quo vadis: Is twitter bitter or are tweets sweet. In: ITNG, pp. 1257–1260 (2010)
5. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: HICSS, pp. 1–10 (2010)
6. Bulmer, M.: Principle of Statistics. Dover Publications, New York (1979)
7. Gruhl, D., Guha, R.V., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: KDD, pp. 78–87 (2005)
8. Katz, E., Lazarsfeld, P.: Personal influence: the part of played by people in the flow of mass communications. Free Press, New York (1955)
9. Kwak, H., Lee, C., Park, H., Moon, S.B.: What is twitter, a social network or a news media? In: WWW, pp. 591–600 (2010)
10. Mishne, G., de Rijke, M.: Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy. In: WWW, pp. 925–926 (2006)
11. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: ACM Conference on Electronic Commerce, pp. 158–167 (2000)
12. Elkin, T.: Just an online minute online forecast, `http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticleartaid=29803`
13. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)

[3] `http://www.amazon.com`

[4] `http://simania.co.il`

[5] `http://news.google.com`